Reuse of search experience for resource transformation

Conference or Workshop Item

Reuse of Search Experience for Resource Transformation

Peter Milne[1], Nirmalie Wiratunga[1], Robert Lothian[1], and Dawei Song[1]

School of Computing
The Robert Gordon University
Aberdeen AB25 1HG, Scotland, UK
{p.m.milne,n.wiratunga,r.m.lothian,d.song}@rgu.ac.uk
http://www.comp.rgu.ac.uk

Abstract. The advent of Web 2.0 has created a proliferation of resource sharing sites where individual users tag resources. Retrieval performance is good when users share the same vocabulary, but deteriorates when users have diverging vocabularies. In this paper we propose a novel method of reusing search experience to transform the underlying representation of tagged resources. The aim is to favour those tags that best correspond to community consensus. A CBR approach is presented to learn from user search histories, modifying resource tags in response to implicit user feedback. We evaluate this method on a prototype image retrieval system iFETCH. Our evaluation shows that resource transformation progressively increases the ranking of those images that are generally deemed relevant by similar search sessions. Our results also confirm that the casebase weight update mechanism is more robust to erroneous user feedback compared to a naive constant weight update strategy.

1 Introduction

Social searching and browsing have in recent years become increasingly popular with the advent of Web2.0 applications. These applications allow users to share resources such as documents, images, videos, music and Blogs on the Web. Resource annotation in the form of tagging is freely used and refers to the free association of keywords to resources by members of a given community using their own vocabulary [12]. The term folksonomy is now commonly used to refer to the bottom up structures that emerge as a result of such social tagging [1]. In recent years, social tagging has become popular, with Wikipedia, Flikr and del.icio.us websites a testament to this trend.

The absence of textual content presents a significant challenge for index creation for multimedia retrieval [8]. However in recent years social tagging has presented itself as a useful indexing knowledge source.An interesting emergent problem is the absence of a controlled tag vocabulary. Although this is attractive for users it is obviously not ideal for vocabulary management and multimedia content comparison. In this paper we investigate how to utilise user search and browsing experiences to evolve indexing vocabularies so as to capture consensus. We achieve this by transforming resource representations by altering individual tag weights thereby moving the resource closer to the queries that are commonly used to search that resource.

Resource transformation involves the modification of the underlying representation of a given resource [9]. Unlike feature weighting here feature values are incremented or alternatively decremented. Identifying which feature values to update and by how much are key concerns for transformation. Here we rely on user interaction and implicit relevance judgments. We introduce a casebased tag weight update mechanism by capturing previous user search experiences. Local neighbourhood similarity values are used to control the amount by which a weight is updated. We demonstrate our approach on an image retrieval task. Initial evaluation results are promising, showing that the technique increases the ranking of those images that the majority of users have deemed as relevant. Our results also show that the casebase weight update mechanism is more robust in the presence of atypical users.

In Section 2 we discuss the role of CBR within Web2.0 applications. Resource transformation using a casebased approach appears in Section 3 followed by a description of the image retrieval prototype, iFETCH, in Section 4. Our experimental design and initial results are discussed in Section 5. We conclude in Section 7 after presenting related work in Section 6.

2 Learning from Searching and Browsing Experiences

With information retrieval (IR) systems, users initiate retrieval through keyword-based search queries. Standard meta-search engines typically return a set of resources ranked according to their relevance to the query. A resource for example can be a document, image, video or Blog; and relevance typically estimates the level of term overlap in textual content. Like CBR, resource representation, indexing and similarity measures are key IR system design considerations. The vector space model is typically used for resource representation combined with an inverted list for indexing and the cosine similarity metric for retrieval [2]. Each dimension in a vector representation is a different term, where a non-zero value indicates the presence of that term. Term importance or term weighting within each resource is typically captured as a function of term frequency and is known to improve search retrieval and ranking.

Tagging provides an additional source of knowledge for the standard feature vector representation. In fact studies suggest a greater overlap between the tag-query vocabularies when compared to the content-query vocabularies [11]. As a result it is increasingly common to represent resources based on just the tagging vocabulary. However the notion of tag importance within each resource is hard to capture, this is because a tag is simply assigned to a resource and is either present or absent. The situation is further exacerbated by the absence of a controlled tagging vocabulary, particularly with broad folksonomies where users can tag a resource with any set of tags that they see fit.

The question then is what knowledge can we use to establish within resource tag importance? User consensus on tags (tag popularity) and user selection patterns are both useful for this purpose. Unlike tag popularity which is easily captured by means of tag counters, user selection patterns require more sophisticated capture and integration mechanisms. For instance, if a user selects a resource as relevant to a query, then the query terms that are also used to tag this resource must be deemed important by this

user. This allows us to implicitly identify which tag weights should be increased or alternatively decreased. However, some users might mistakenly select a resource as relevant or be atypical and diverge from the majority. Therefore, instead of considering single user behaviour in isolation, it is far more useful to rely on collective user selection patterns. We need to also consider how much weights should be altered after each user session? The answer depends very much on the level of similarity in user selection patterns. To this end we exploit search history information to learn how much to update a tag weight by. We propose to maintain a casebase of previous queries, together with user selected resources. The more similar the cases to the recent user session the more reinforcement for tag weight update.

3 Case Representation and Transformation

A case consist of a query and a set of resources judged as relevant to the query by the user. The reliability of a user's feedback is a function of the similarity between the current search and the most similar cases in the casebase. These cases represent previous searches where the query and the relevant resources are the most similar to the current search. Each tag associated to a resource is given a weight proportional to how representative this tag is for this resource. This weight is consistently updated as the casebase evolves and more information is deduced from previous searches.

3.1 Case Representation

A case is a pair (Q, RR), where Q is a query and RR is a set of resources judged relevant to this query. A query Q is represented by a set $\{q_1, \ldots, q_l\}$ of keywords q_i present in the query. A set of relevant resources RR is represented by a set $\{R_1, \ldots, R_m\}$ of resources R_j judged by a user as relevant to the query Q. Each resource R_j is itself represented by a set $\{t_1, \ldots, t_n\}$ of tags t_k, where each tag has an associated weight w_k. A resource is normalised so that the weights of all associated tags t_k sum up to 1. In order to ascertain and utilise collective user judgments, the similarity between 2 cases is computed by aggregating the similarity between the query and the similarity between the set of relevant resources. The more similar the cases, the more similar the user search experiences.

3.2 Resource Transformation

The purpose of maintaining a casebase of previous searches and computing the similarities to the current search keywords is to optimise the tag weights associated to each resource. We propose to penalise tags associated to resources selected as relevant if they are not in the query. The weight reduction of such tags occurs as a side effect of the weight renormalisation stage. The weight update is a function of the similarity between the current search, or test case, and previous searches, or training cases. The more similar the test case is to the nearest cases in the casebase, the more the current

users's feedback is aligned with previous users' feedback and therefore deemed as more reliable. The tags weights of a resource $R \in RR$ are updated as follows:

$$w_i{}^{temp} = w_i * update(R, Q)$$
$$update(R, Q) = \begin{cases} 1 & \text{if } t_i \in Q \\ f(CB, Q) & \text{if } t_i \notin Q \end{cases}$$

$$f(CB, Q) = (1 - \overline{Sim}_k(CB, Q))^{\alpha}$$

Here $\overline{Sim_k}$ is the average similarity between Q and its top k neighbours in CB. We use the cosine similarity metric and α is a parameter controlling the severity of the penalty. After each run, a resource's representation is updated so that the weights of each resource tag not present in the query is reduced relatively to the similarity between the current query and the most similar cases in our casebase. In turn, weight renormalisation will have the effect of increasing the weights of each resource tag present in the query. The renormalisation is achieved as follows:

$$w_i{}^{new} = \frac{w_i{}^{temp}}{\sum_{j=1}^{n} w_j{}^{temp}}$$

The overall impact of updating weights is to increase the importance of commonly used tags in the similarity metric used during retrieval, therefore achieving a ranking of resources more aligned with a global consensus of opinion.

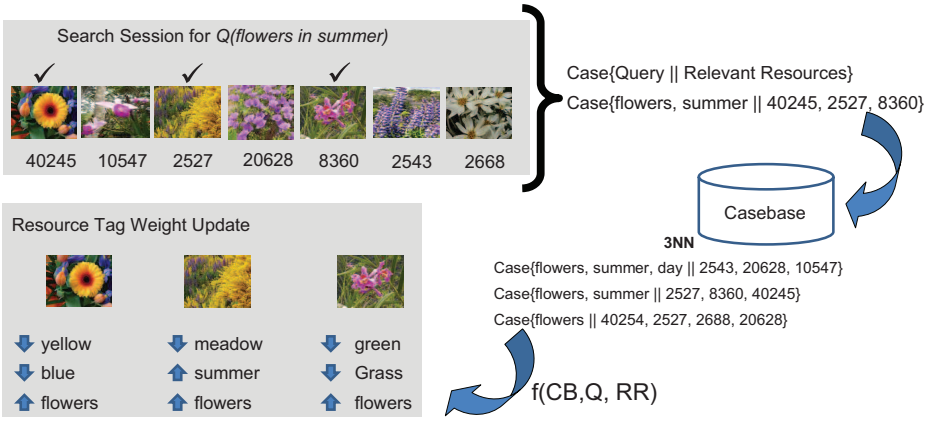

Fig. 1. Tag weights updated using a casebase

Figure 1 illustrates an example of a tag weight update session for images. The user executes a search with keywords *flowers in summer*. From the search results the users selects 3 images 40245, 2527, 8360 that they feel are relevant to their information need. A new case made up of both the keywords and the identifiers for the images selected

is then created and presented to the casebase. The k nearest neighbours are identified by calculating the cosine similarity between the new case and those in the casebase. A function of this similarity is then used to update the weights of each of the tags, where tags that were not in the query have their weights decreased, and tags common to the query have their weights increased. Finally tag weights of each of the selected resources are updated before the new case is added to the casebase for future searches.

4 The iFETCH Prototype

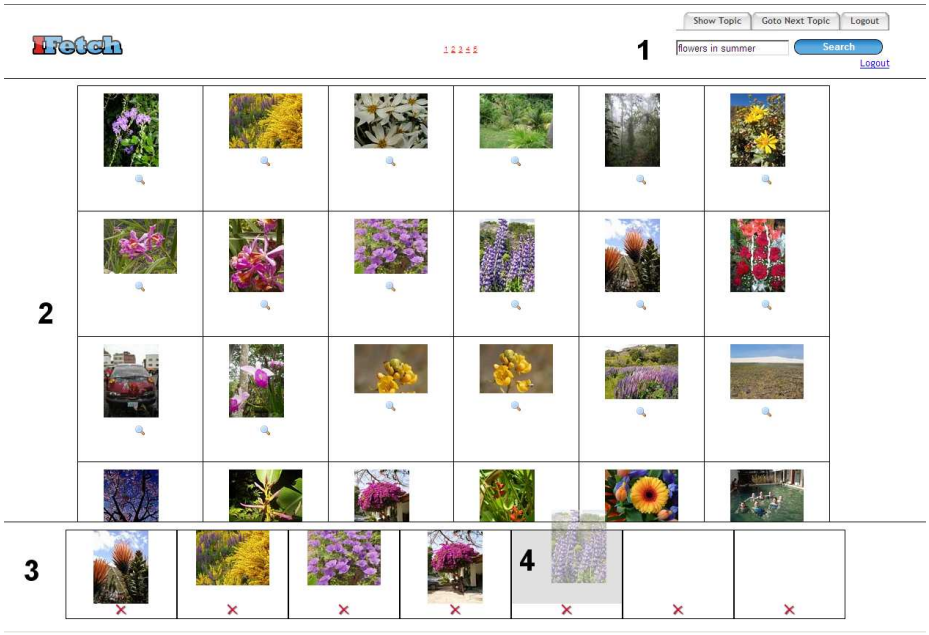

Fig. 2. The iFETCH user interface

To test our resource transformation method a demonstrator Web2.0 application, iFETCH was developed for image retrieval. Figure 2 shows the main components of the user interface. The interface header panel contains the search box to the top right (area 1) allowing users to input their search terms. Also in the centre of the header panel is the page selectors which are shown when there are multiple pages of results returned. Search result appear in the center (in area 2). Each image in this panel can be enlarged by clicking a magnifying glass icon below the image. Relevant images can be dragged into the task panel (shown by area 3) at the bottom of the interface. Each of the images in the task panel can be swapped back and forth or removed using a cross icon below each image. This is illustrated in the figure with an image being dragged by a user into the task panel (in area 4). For demonstration purpose the task panel currently allows a

selection of upto seven images. This is in keeping with Miller's [15] findings that the usual limit of the human short-term memory store is around seven units. IFETCH also allows users to execute multiple searches for the same information need while retaining any previously selected relevant images in the task panel. This is to allow the user to combine selections from multiple searches for the same information need. This is maintained as part of the user's profile, allowing them to amend previous queries and track their history of information needs.

5 Experimental Evaluation

We evaluate IFETCH incorporating the casebased weight update mechanism with a constant update function without a casebase. We also investigate the robustness of retrieval in the presence of noise. By noise we mean user's that may incorrectly identify resources as relevant or deviate from majority consensus. Precision at 10 is a good measure for evaluating the performance of web based retrieval systems [14]. This is due to the fact that the majority of people will only examine the first page of 10 results. Mean Average Precision (MAP) is a more recent metric that calculates the precision after each relevant document's retrieved rank [5] . These scores are then averaged over the number of relevant documents retrieved. For instance, if a run retrieves 3 relevant documents, one ranked 3rd, one ranked 5th and one ranked 10th. The precision at rank 3, 5 and 10 is calculated and averaged. In our evaluation, we used both metrics to compare the 2 systems: Precision at 10 (P10) and Mean Average Precision (MAP).

It is generally accepted that evaluation is a challenge when user interaction is central to the techniques being evaluated. This is mainly due to the cost involved with large scale user trials, requiring live user participation. In order to get round this problem we developed a test harness to simulate user query generation and resource selection. Instead of collecting actual data from user trials we simulate search sessions using the IMAGECLEF'06 collection containing a set of images and their relevance to 60 topics.

5.1 Query Generation and User Trial Simulation

The IMAGECLEF 2006 dataset contains 20,000 images from many locations around the world. The majority of the images provided in the dataset are images from an independent travel company organising adventure and language trips to South America. The images within the dataset are varied and are realistic in terms of what we would expect to search in a web based image retrieval system. Figure 3 shows an example of a typical topic (animal swimming) and an associated image annotation. Topic annotations capture typical search needs and have been generated from search logs with the aim to provide a balanced and representative set of information needs. Importantly IMAGECLEF also has images labelled as relevant for each topic. We assume that these ground-truth images would be the ones that typical users might select for a given query (for a specific topic). As such the retrieval rank of ground-truth images provides the basis for comparing retrieval performance between user trials. The more ground truths ranked at the top the better the retrieval.

<TOPIC>
<NUM> Number: 5 </NUM>
<TITLE> animal swimming </TITLE>
<NARR>
Relevant images will show one or more animals (fish, birds, reptiles, etc.) swimming in a body of water.

Images of people swimming in water are not relevant.

Images of animals that are not swimming are not relevant.
</NARR>

</TOPIC>

<IMAGE>
<NO>annotations/34/34160.eng</NO>
<TITLE>Snowboarder sitting in the snow</TITLE>
<DESCRIPTION>a snowboarder with a brown jacket,
black trousers and a black snowboard is sitting in the snow;
</DESCRIPTION>
<NOTES></NOTES>
<LOCATION>Mondsee, Austria</LOCATION>
<DATE>December 2003</DATE>
<IMAGE>images/34/34160.jpg</IMAGE>
<THUMBNAIL>thumbnails/34/34160.jpg</THUMBNAIL>
</IMAGE>

Fig. 3. The IMAGECLEF example topic and image annotation

Each image is allocated tags by extracting stemmed keywords from its description (see Figure 3). Tag weights are initialised and length normalised. We simulate a search task as selecting up to 7 images from those retrieved for a system generated query. Queries are generated by extracting keywords from a topic's title and as such play the role of a user query. For each query we then simulate 105 user trials, whereby 7 images are randomly selected as relevant from the retrieved set in a given user trial. A perfect user is simulated by randomly selecting ground-truth images from the set of retrieved images whilst a random selection forming a mix of ground truths and other irrelevant images simulates an imperfect user. In this way the level of error is managed by using a mixing parameter. A case is created for each simulated user trial and tag weights accordingly updated at each trial.

5.2 Results

Figure 4 illustrates the evolution of MAP and P10 results averaged over the 60 topics for each ordered trial. During the first 25 runs, we assume a perfect user. The following 15 runs simulate unreliable user feedback. This is simulated by randomly selecting only 20% of ground-truth images and 80% of irrelevant images. This pattern is repeated for the following 65 runs.

The first trial indicates the MAP and P10 result when no resource transformation is used. As expected, resource transformation significantly improves upon this first trial in consequent trials. Both MAP and P10 results peek faster when no casebase is used initially. This is because, during the 25 first runs, ground-truth images are selected and the relevant tag weights of each of these images are promoted after every run. However, with a casebase, the same image has to be selected multiple times before the similarity between the new case and cases in the casebase is sufficient to significantly promote tag weights. This initial phase can be likened to a standard CBR system's performance during the case generation or authoring phase.

The advantage of the matured casebase becomes more apparent with the introduction of user error. This can be seen with the introduction of erroneous user feedback between trials 25 to 40, where MAP drops considerably when not using a casebase,

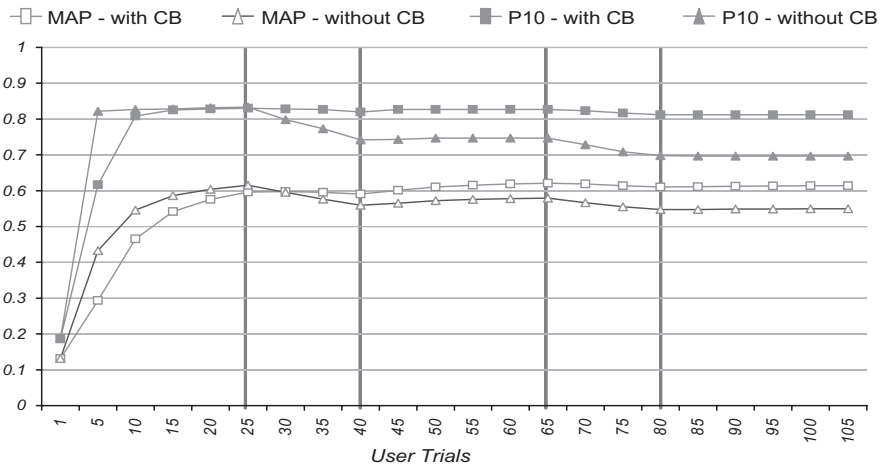

Fig. 4. Effect of erroneous feedback on MAP and P10

whilst remaining relatively stable when using the casebase. This is because, with a casebase, an irrelevant image has to be selected multiple times before it can be promoted to a higher rank i.e. the majority of users must agree that these images fulfill the information need better than those previously selected. In the absence of a casebase, an irrelevant image gets promoted quickly to a higher rank as soon as it gets selected. This phenomenon is illustrated even more clearly in Figure 4, with the P10 graphs. Without the use of a casebase, P10 suffers a drop at each introduction of erroneous feedback and does not recover, even when consistent relevant feedback follows. It is also important to note that the average P10 achieved over the 60 topics reaches a level of over 0.8. This is much higher than the average MAP. While MAP illustrates the spread of relevant images over the retrieval, P10 illustrates the amount of relevant images ranked in the top 10. This is a more relevant result as, in a real web based system, users would tend to look at the top images within the first page and are less likely to browse other pages.

6 Related Work

Many studies have been conducted into user tagging behaviour with a view to identify how best to utilise tags as meta-knowledge sources for collection organisation and searching [12, 13]. Although the absence of any controlled vocabulary is viewed as a drawback, studies have shown that frequently used tags are still a good indicator of user query terms [11]. In our work we leverage on user queries as a means to refine the explicit association of tags to images. A casebase is used to capture tag popularity and social consensus about tags. Case similarity values are then used to implicitly refine the tag-to-image associations. The approach is sufficiently generic and can be transfered to other resources commonly encountered within folksonomies.

Refining associations between tags and their resources can be viewed as a form of tag recommendation but also as a form of relevance feedback. In traditional IR rel-

evance feedback is used to refine a query entered by the user. The popular Rocchio algorithm uses feedback from a user's explicit preferences to form two centroids; representative of relevant and irrelevant resources [6]. Using a linear combination query terms are then promoted if they appear in the relevant centroid and demoted otherwise. Here promotion can also be viewed as query expansion. Reliance on users for explicit feedback is a drawback which has been addressed by research into implicit feedback capture. Typically click-throughs and mouse moves are used for this purpose [7]. An interesting aspect here is o facilitate unobtrusive and effective feedback capture though clever interface design [8]. iFETCH also adopts a similar emphasise whereby standard areas for query entry and ranking are augmented with additional drag-and-drop work areas to allow for task management.

Document transformation work in machine learning, aims to refine a document's representation by learning from queries that led to its access. Unlike explicit and implicit relevance feedback mechanisms from IR, here querying experiences are utilised as a means to assist future user's with similar search needs. The idea is to reuse query and relevance feedback knowledge to improve descriptions of selected resources. In [9] indexing descriptors of documents on the web are modified, by incrementally updating it to better match each query that is used to retrieve it. However to directly apply such an approach within a folksonomy setting would be naive. This is because one needs to influence the update according to consensus and not simply on the basis of individual user queries. We achieve this by using an update function that is directly influenced by similarity to previous user search sessions i.e. their queries and chosen relevant resource sets. The more consensus amongst users about a query and its relevant resource set the higher the average similarity leading to higher promotion and demotion of tags.

Use of consensus within a community of like-minded people has been applied to improve document ranking in [10]. A separate representation of documents is maintained so as to to capture document relevance preferences for similar queries entered by users within a community. A new representation for selected documents is generated from snippet text that is returned by standard meta-search engines such as Google or Yahoo. This approach has the useful property of altering surrogate representations of documents instead of the original document representations themselves. As such one can imagine the maintenance of multiple community views through multiple surrogate representations. One drawback here is its reliance on textual snippets and as such does not easily lend itself to other forms of resources within folksonomies: such as images and videos. However the general idea of maintaining multiple community views is still very interesting and an area we intend to explore in the future.

7 Conclusion and Future Work

This paper presents and initial approach to transforming resource representations by altering the weights of tags associated to resources. We present a novel casebased approach to help control the amount by which weights are updated. The casebase maintains previous search experiences consisting of query and relevance judgments. A small-scale evaluation of the iFETCH demo system shows that resource transformation to result in far superior retrieval performance as it learns from each user session. Further-

more the casebase weight update approach remains resistant to user error, and outperforms a constant weight update approach.

The simulation of query generation and user trials in our evaluation is particularly novel. In future work it would be interesting to include new features to this evaluation methodology, for example, to allow variation in queries within topics, thus simulating user's differing vocabularies. Another interesting future direction would be the inclusion of multiple representations for groups of users, transforming these representations based on the group's vocabulary and the behavior of individual's within that group. We also intend to evaluate our weight update strategies using real users with the IFETCH prototype to ascertain if the technique works in a real life scenario.

References

1. Jaschke, R, Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G.: Tag recommendations in folksonomies. *Proceedings Knowledge Discovery in DBs* 506–514, Springer (2007)
2. Salton. G., Wong, A., Yang, C.: A vector space model for automatic indexing *Communications of the ACM* 18(11) 613–620, (1975)
3. Hotho, A., Jschke, J., Schmitz, C., Stumme, G.: FolkRank: A Ranking Algorithm for Folksonomies. In *Proceedings of Fachgruppe Information Retrieval (FGIR)* (2006)
4. Jansen, B. J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management* 36 207–227 (2000)
5. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International Special Interest Group in Information Retrieval (SIGIR)* 33-40 ACM Press (2000)
6. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science* 41(4) 288–297 (1990)
7. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback *Proceedings of the 28th Annual International Special Interest Group in Information Retrieval (SIGIR)* 154–161 ACM Press (2005)
8. White , R., Ruthven, I., Jose, J.: An implicit feedback approach for interactive information retrieval *Information Processing and Management* 42(1) 166–190 (2006)
9. Kemp, C., Ramamohanarao, K.: Long-Term Learning for Web Search Engines, *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)* 263–274 Springer (2002)
10. Boydell, O., Smyth, B.: Enhancing Case-Based, Collaborative Web Search. *Proceedings of the 7th International Conference on Case-Based Reasoning* 329–343 Springer (2007)
11. Suchanek, F., Vojnovic, M., Gunawardena, D.: Social tags: meaning and suggestions. *Proceedings of the 17th Conference on Information and Knowledge Management (CIKM)* 223–232 ACM Press (2008)
12. Van Damme,C., Hepp,M., Coenen, T.: Quality Metrics for Tags of Broad Folksonomies. *Proceedings of International Conference on Semantic Systems (I-SEMANTICS)* 118–125 (2008)
13. Farooq, U., Kannampallil, T., Song, Y. Ganoe, C., Carroll, J. Giles, L.: Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics *Proceedings of International Conference on Supporting Group Work (GROUP)* 351–360 ACM Press (2007)
14. Anh, V. N., Moffat, A.: Improved retrieval effectiveness through impact transformation. *Australian Computer Science Communications* 24(2) 41–47 (2002)
15. G., A., Miller.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review* 63 81–97 (1956)